Chương 8

Các thuật toán chuẩn hóa cơ sở dữ liệu

Nội dung trình bày

- Bao đóng của tập phụ thuộc hàm
- Bao đóng của tập thuộc tính
- Tập PTH tương đương và tối thiểu
- Thuật toán xác định khóa
- Các thuật toán chuẩn hóa cơ sở dữ liệu

Bao đóng của tập PTH

- Ta gọi f là một phụ thuộc hàm được suy dẫn từ F nếu tồn tại một chuỗi phụ thuộc hàm: $f_1, f_2, \ldots, f_n$ sao cho $f_n = f$ và mỗi f_i là một thành viên của F hay được suy diễn từ những phụ thuộc hàm j = 1, …, i-1 trước đó nhờ vào luật dẫn.
- Bao đóng của F, ký hiệu F^+, là tập hợp các phụ thuộc hàm được suy diễn từ F nhờ vào hệ tiên đề Armstrong.
  - $F \subseteq F^+$.
  - F gọi là đầy đủ nếu $F = F^+$.

Bao đóng của tập thuộc tính

- Bao đóng của tập thuộc tính X xác định trên tập phụ thuộc hàm F ký hiệu là X_F^+ là tập hợp tất cả các thuộc tính có thể suy ra từ X.
  - $X_F^+ = \{A \in U \mid X \rightarrow A \in F^+\}$
- Nhận xét
  - Bài toán xác định một phụ thuộc hàm có được suy diễn từ tập phụ thuộc hàm cho trước?
    - $X \rightarrow Y \in F^+ \Leftrightarrow Y \subseteq X_F^+$.

Thuật toán tìm bao đóng của tập thuộc tính

- Nhập: U, F và $X \subseteq U$
- Xuất: X_F^+
- Thuật toán 7.1
  - $X^+ = X$;
  - Repeat
    oldX$^+$ = X$^+$;
    For (mỗi phụ thuộc hàm $Y \rightarrow Z \in F$) do
      if ($Y \subseteq X^+$ và $Z \not\subset X^+$) then $X^+ := X^+ \cup Z$;
  - Until (oldX$^+$ = X$^+$).

Ví dụ tìm bao đóng X$^+$

- Cho:
  - F = {AB $\rightarrow$ C, BC $\rightarrow$ D, D $\rightarrow$ EG}.
  - X = BD.
- Tính X$^+$:
  - X$^+$ = BD.
  - Lặp 1:
    - oldX$^+$ = X$^+$ = BD
    - D $\rightarrow$ EG, thêm EG vào X$^+$ ta được X$^+$ = BDEG.
  - Lặp 2:
    - oldX$^+$ = X$^+$ = BDEG
    - Không thể mở rộng thêm X$^+$.
  - Vậy X$^+$ = BDEG.

Bài toán xác định phụ thuộc hàm suy diễn

- Cho F = {AB $\rightarrow$ C, A $\rightarrow$ D, D $\rightarrow$ E, AC $\rightarrow$ B}
- Hai phụ thuộc hàm AB $\rightarrow$ E và D $\rightarrow$ C có được suy diễn từ F hay không?

X	X_F^+
AB	ABCDE
D	DE

AB $\rightarrow$ E Được suy diễn từ F

Tập phụ thuộc hàm tương đương

- Tập phụ thuộc hàm F được nói là phủ tập phụ thuộc hàm G nếu tất cả các phụ thuộc hàm trong G có thể được suy diễn từ F.
  - F phủ G $\Leftrightarrow$ G $\subset$ F^+.
  - $\forall X \rightarrow Y \in G$, nếu $Y \subseteq X_F^+$ thì F phủ G.
- Hai tập phụ thuộc hàm F và G là tương đương nếu tất cả các phụ thuộc hàm trong F có thể được suy diễn từ G, và tất cả các phụ thuộc hàm trong G có thể được suy diễn từ F.
  - F và G tương đương $\Leftrightarrow$ $F^+ = G^+$.
  - F và G là tương đương $\Leftrightarrow$ F phủ G và G phủ F.

Tập phụ thuộc hàm tối thiểu

- Vấn đề thừa phụ thuộc hàm
  - {A → B, B → C, A → C}, vì A → C được suy diễn từ {A → B, B → C}
    A → B, B → C ⇒ A → C (luật bắc cầu).
- Vấn đề thừa thuộc tính
  - {A → B, B → C, A → CD}, vì A → CD được suy diễn từ {A → B, B → C, A → D}
    A → B, B → C ⇒ A → C (luật bắc cầu)
    A → C, A → D ⇒ A → CD (luật hợp).
  - {A → B, B → C, AC → D}, vì AC → D được suy diễn từ {A → B, B → C, A → D}
    A → B, A → D ⇒ A → BD (luật hợp)
    A → BD ⇒ AC → BCD (luật tăng trưởng)
    AC → BCD ⇒ AC → D (luật phân rã).

Tập phụ thuộc hàm tối thiểu

- Tập phụ thuộc hàm F là tối thiểu nếu nó thỏa các điều kiện sau:
  - Mọi phụ thuộc hàm của F chỉ có một thuộc tính ở vế phải.
  - Không thể thay thế bất kỳ phụ thuộc hàm X → A nào trong F bằng phụ thuộc hàm Y → A, với Y ⊂ X mà vẫn có được một tập phụ thuộc hàm tương đương với F (tức là, không có thuộc tính dư thừa trong phụ thuộc hàm)
  - Không thể bỏ đi bất kỳ một phụ thuộc hàm nào trong F mà vẫn có được một tập phụ thuộc hàm tương đương với F (tức là, không có phụ thuộc hàm dư thừa).
- Phủ tối thiểu của tập phụ thuộc hàm F là tập phụ thuộc hàm tối thiểu tương đương với F.
  - Tất cả các tập phụ thuộc hàm đều có ít nhất một phủ tối thiểu.

Thuật toán tìm phủ tối thiểu

- Nhập: tập phụ thuộc hàm F.
- Xuất: phủ tối thiểu F.
- Thuật toán 7.2
  - *B1*: G := ∅.
  - *B2*: Với mỗi $X \to Y \in F$, $Y = \{A_1, \ldots, A_k\}$, $A_i \in U$
    $G := G \cup \{X \to A_i\}$.
  - *B3*: (Loại bỏ thuộc tính thừa trong các phụ thuộc hàm)
    Với mỗi $X \to A \in G$, với mỗi thuộc tính $B \in X$ nếu $A \in (X - B)_G^+$ thì
    $G := (G - \{X \to A\}) \cup \{(X - B) \to A\}$.
  - *B4*: (Loại bỏ phụ thuộc hàm thừa)
    Với mỗi $X \to A \in G$
    $G' := G - \{X \to A\}$
    Nếu $A \in X_{G'}^+$ thì $G := G - \{X \to \{A\}\}$.

Ví dụ tìm phủ tối thiểu

- Tìm phủ tối thiểu của $F = \{A \to BC, A \to B, B \to C, AB \to C\}$
  - *B1*: G = ∅.
  - *B2*: $G = \{A \to B, A \to C, B \to C, AB \to C\}$.
  - *B3*: Xét $AB \to C$
    $(B)_G^+ = BC$
    $G = \{A \to B, A \to C, B \to C\}$.
  - *B4*: $A \to C$ thừa.
    $G = \{A \to B, B \to C\}$.

Thuật toán tìm khóa

- Nhập: R(U) và tập phụ thuộc hàm F.
- Xuất: khóa K của R.
- Thuật toán 7.3.1
  - *B1*:

    K = U = $\{A_1, \ldots, A_n\}$;

    i = 1;
  - *B2*:

    Nếu $U \subseteq (K - A_i)_F^+$ thì $K = K - A_i$.

    i = i + 1;

    Nếu i > n thì sang *B3*. Ngược lại, tiếp tục *B2*.
  - *B3*:

    Xuất K.

Ví dụ tìm khóa

- Cho R(U), U = {A, B, C, D, E, F, G}.
  - F = {B $\rightarrow$ A, D $\rightarrow$ C, D $\rightarrow$ BE, DF $\rightarrow$ G}.
- Tìm khóa của R
  - B1:

    K = ABCDEFG.
  - B2:
    - Lặp 1: $(BCDEFG)_F^+ = BCDEFGA \Rightarrow K = BCDEFG$.
    - Lặp 2: $(CDEFG)_F^+ = CDEFGBA \Rightarrow K = CDEFG$.
    - Lặp 3: $(DEFG)_F^+ = DEFGCBA \Rightarrow K = DEFG$.
    - Lặp 4: $(EFG)_F^+ = EFG$.
    - Lặp 5: $(DFG)_F^+ = DFGCBEA \Rightarrow K = DFG$.
    - Lặp 6: $(DG)_F^+ = DGCBEA$.
    - Lặp 7: $(DF)_F^+ = DFCBEAG \Rightarrow K = DF$.
  - B3:

    Khóa là K = DF.

Thuật toán tìm tất cả khóa

- Nhập: tập PTH F xác định trên lược đồ R(U).
- Xuất: tất cả khóa của R.
- Thuật toán 7.3.2
  - *B1*:
    Xây dựng $2^n - 1$ tập con khác rỗng của $U = \{A_1, \ldots, A_n\}$;
    $S = \{\}$;
  - *B2*:
    Với mỗi tập con $\varnothing \neq X \subseteq U$
      Tính bao đóng X_F^+;
      Nếu $U \subseteq X_F^+$ thì $S = S \cup \{X\}$.
  - *B3*:
    $\forall K_1, K_2 \in S$, nếu $K_1 \subset K_2$ thì $S = S - K_2$.
  - *B4*:
    S là tập các khóa của R.

Ví dụ tìm tất cả khóa

- Cho R(U), U = {A, B, C, D, E, F}.
  - F = {AE $\rightarrow$ C, CF $\rightarrow$ A, BD $\rightarrow$ F, AF $\rightarrow$ E}.
- Tìm tất cả khóa của R

X	X_F^+	Siêu khóa	Khóa
A	A		
B	B		
…			
AE	AEC		
…			
ABD	ABDFEC	ABD	ABD
…			
BCD	BCDFAE	BCD	BCD
…			
ABCD	ABCDEF	ABCD	
…			

Tập siêu khóa S = {ABD, BCD, ABCD, ABDE, BCDE, ABCDE, ABDF, BCDF, ABCDF, ABDEF, BCDEF, ABCDEF}

Thuật toán tìm tất cả các khóa cải tiến

- Tập hợp các thuộc tính ở vế trái các PTH của F : VT_F
- Tập hợp các thuộc tính ở vế phải các PTH của F : VP_F
- Tập nguồn là tập các thuộc tính không xuất hiện ở vế phải các PTH của F : $TN = U - VP_F$
- Tập đích là tập các thuộc tính có xuất hiện ở vế phải và không xuất hiện ở vế trái các PTH của F : $TD = U - (TN \cup VT_F)$
- Tập trung gian là tập các thuộc tính vừa xuất hiện trong vế trái, vừa xuất hiện trong vế phải của các PTH của F : $L = U - (TN \cup TD)$
- **Nếu K là khóa của R thì tồn tại $L_i \subseteq L$ sao cho $K = TN \cup L_i$.**

Ví dụ tìm tất cả khóa cải tiến

- VT_F = {A, B, C, D, E, F}
- VP_F = {A, C, E, F}
- $TN = U - VP_F$ = {B, D}
- $TD = U - (TN \cup VT_F) = \varnothing$
- $L = U - (TN \cup TD)$ = {A, C, E, F}

L_i	$Li \cup TN$	$(Li \cup TN)_F^+$	Siêu khóa	Khóa
∅	BD	BDF		
A	ABD	ABDFEC	ABD	ABD
C	CBD	CBDFAE	CBD	CBD
E	EBD	EBDF		
F	FBD	FBD		
AC	ACBD	ABCDEF	ABCD	
…				

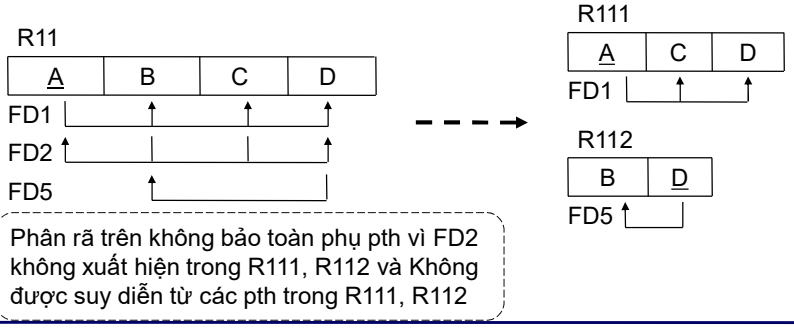
Phân rã quan hệ

Cho lược đồ quan hệ phổ quát R với tập hợp tất cả các thuộc tính của cơ sở dữ liệu $\{A_1, \ldots, A_n\}$ và tập phụ thuộc hàm F xác định trên các thuộc tính của R.

- Phân rã quan hệ là quá trình sử dụng các thuật toán chuẩn hóa để tách quan hệ phổ quát R thành một tập hợp các quan hệ $D = \{R_1, \ldots, R_m\}$. D được gọi là một phân rã của R.
- Một phân rã phải đảm bảo:
  - Bảo toàn thuộc tính nghĩa là mỗi thuộc tính của R phải xuất hiện trong ít nhất một R_i.
  - Các quan hệ R_i phải ở dạng chuẩn 3NF hoặc BCNF.

Phân rã bảo toàn phụ thuộc hàm

- Cho quan hệ R(U) và tập pth F. Giả sử $D = \{R_1, \ldots, R_m\}$ là một phân rã của R.
  - Phép chiếu của F trên R_i là $\pi_{Ri}(F) = \{X \rightarrow Y \in F^+ : X \cup Y \subset R_i\}$.
  - D được gọi là phân rã bảo toàn phụ thuộc hàm đối với F nếu $(\pi_{R1}(F) \cup \ldots \cup \pi_{Rm}(F))^+ = F^+$. Nghĩa là các pth trong F hoặc xuất hiện trực tiếp trong các quan hệ R_i hoặc được suy diễn từ các pth trong R_i.



Phân rã trên không bảo toàn phụ pth vì FD2 không xuất hiện trong R111, R112 và Không được suy diễn từ các pth trong R111, R112

Phân rã bảo toàn phụ thuộc hàm

R11	A	B	C	D
	1	α	β	2
	2	β	γ	3
	3	α	δ	2

R111	A	C	D
	1	β	2
	2	γ	3
	3	δ	2
	4	β	4

R112	D	B
	2	α
	3	β
	4	α

A	B	C	D
1	α	β	2
…	…	…	…
4	α	β	4

Thêm bộ (4, β, 4) vào R111 và (4, α) vào R112 thì trạng thái csdl sẽ không thỏa PTH FD2

Thuật toán phân rã bảo toàn PTH

- Định lý 7.1

  Luôn tồn tại một phân rã bảo toàn pth D = {R_1, …, R_m} đối với tập phụ thuộc hàm F của một quan hệ phổ quát R sao cho các R_i ở dạng chuẩn 3NF.

- Thuật toán 7.4
  - Nhập: R(U), U = {A_1, …, A_n} và tập pth F.
  - Xuất: Phân rã D = {R_1, …, R_m}, R_i ở dạng chuẩn 3NF.
  - *B1*:

    Tìm phủ tối thiểu G của F.
  - *B2*:

    Với mỗi $X \rightarrow A_i \in G$, xây dựng quan hệ $R_i(U_i)$, $U_i = X \cup \{A_j\}$. Khóa chính của R_i là X.

Thuật toán phân rã bảo toàn PTH

- *B3*:

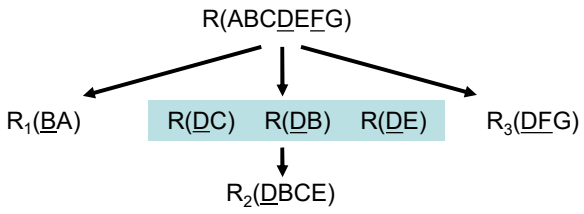
  Giả sử xong B2 ta có các quan hệ R_1, …, R_m. Nếu ($U_1 \cup \ldots \cup U_m$) $\neq$ U thì xây dựng thêm lược đồ $R_{m+1}(U_{m+1})$, $U_{m+1} = U - (U_1 \cup \ldots \cup U_m)$. Khóa chính của R_{m+1} là U_{m+1}.
- *B4*:

  Phân rã D là tập hợp các quan hệ R_i.

Ví dụ phân rã bảo toàn PTH

- Cho R(ABCDEFG), F = {B $\rightarrow$ A, D $\rightarrow$ C, D $\rightarrow$ EB, DF $\rightarrow$ G}
- Tìm phân rã bảo toàn pth của R
  - *Tìm phủ tối thiểu*

    G = {B $\rightarrow$ A, D $\rightarrow$ C, D $\rightarrow$ B, D $\rightarrow$ E, DF $\rightarrow$ G}.
  - *Cây phân rã*



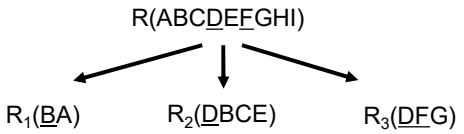
  - *Phân rã bảo toàn pth của R*

    Xuất D = {R_1, R_2, R_3}.

Ví dụ phân rã bảo toàn PTH

- Cho R(ABCDEFGHI), F = {B → A, D → C, D → EB, DF → G}
- Tìm một phân rã bảo toàn pth của R
  - *Tìm phủ tối thiểu*

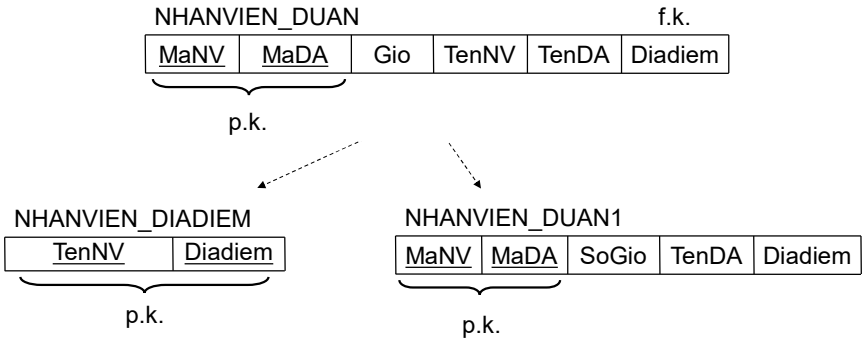
    G = {B → A, D → C, D → B, D → E, DF → G}.
  - *Cây phân rã*



  - Vì $U_1 \cup U_2 \cup U_3$ = {ABCDEFG}, ta đặt R_4(HI).
  - *Phân rã bảo toàn pth của R*

    D = {R_1, R_2, R_3, R_4}.

Phân rã không mất thông tin



NHANVIEN_DUAN

MaNV	MaDA	Gio	TenNV	TenDA	Diadiem
123456789	1	32.5	Hung	San pham X	Tan Binh
123456789	2	7.5	Hung	San pham Y	Thu Duc
333445555	2	10	Nghia	San pham Y	Thu Duc

Phân rã không mất thông tin

NHANVIEN_DIADIEM

TenNV	Diadiem
Hung	Tan Binh
Hung	Thu Duc
Nghia	Thu Duc

NHANVIEN_DUAN1

MaNV	MaDA	SoGio	TenDA	Diadiem
123456789	1	32.5	San pham X	Tan Binh
123456789	2	7.5	San pham Y	Thu Duc
333445555	2	10	San pham Y	Thu Duc

Kết tự nhiên

MaNV	MaDA	Gio	TenDA	Diadiem	TenNV
123456789	1	32.5	San pham X	Tan Binh	Hung
123456789	2	7.5	San pham Y	Thu Duc	Hung
123456789	2	7.5	San pham Y	Thu Duc	Nghia
333445555	2	10	San pham Y	Thu Duc	Hung
333445555	2	10	San pham Y	Thu Duc	Nghia

Phân rã không mất thông tin

- Cho quan hệ R(U) và tập pth F. Giả sử D = {R_1, …, R_m} là một phân rã của R.
  - D được gọi là phân rã không mất thông tin đối với F nếu với mọi trạng thái r ∈ R thì (π_{R1}(r) * … * π_{Rm}(r)) = r.
  - π_{Ri}(r) là phép chiếu trạng thái r lên R_i.
- Định lý 7.2

  Phân rã D = {$R_1(U_1)$, $R_2(U_2)$} của R(U) không mất thông tin đối với tập PTH F nếu và chỉ nếu:
  - $(U_1 \cap U_2) \rightarrow (U_1 - U_2) \in F^+$, hoặc
  - $(U_1 \cap U_2) \rightarrow (U_2 - U_1) \in F^+$.
- Định lý 7.3

  Nếu phân rã D = {R_1, …, R_m} của R không mất thông tin đối với F và phân rã D_i = {Q_1, …, Q_k} của R_i không mất thông tin đối với π_{Ri}(F) thì D' = {R_1, …, R_{i-1}, Q_1, …, Q_k, R_{i+1}, …, R_m} của R cũng không mất thông tin.

Thuật toán phân rã không mất thông tin

- Thuật toán 7.5
  - Nhập: R(U), U = {A_1, …, A_n} và tập PTH F.
  - Xuất: D = {R_1, …, R_m}, R_i ở dạng chuẩn BCNF.
  - *B1*:

    D = {R};
  - *B2*:

    -Nếu có lược đồ $Q(U_Q) \in D$ không ở dạng chuẩn BC thì

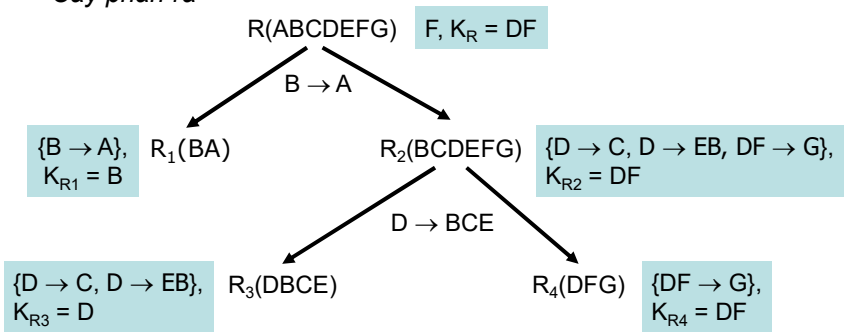
    + Tìm $X \rightarrow Y \in \pi_Q(F)$ làm Q vi phạm điều kiện BC.
    + $D = (D - \{Q\}) \cup Q_1(U_{Q1}) \cup Q_2(U_{Q2})$ với $U_{Q1} = U_Q - Y$ và $U_{Q2} = X \cup Y$.
    + Quay lại B2.

    -Ngược lại, chuyển sang B3.
  - *B3*:

    D là phân rã không mất thông tin của R

Ví dụ phân rã không mất thông tin

- Cho R(ABCDEFG), F = {$B \rightarrow A$, $D \rightarrow C$, $D \rightarrow EB$, $DF \rightarrow G$}
- Tìm một phân rã không mất thông tin của R
  - *Cây phân rã*



  - *Phân rã không mất thông tin của R*

    D = {R_1, R_3, R_4}

Ví dụ phân rã không mất thông tin